Gender Representation in Cinematic Content: A Multimodal Approach

Tanaya Guha Che-Wei Huang Naveen Kumar, Yan Zhu, Shrikanth S. Narayanan
Signal Analysis and Interpretation Lab (SAIL)
University of Southern California, Los Angeles
tanaya@sipi.usc.edu,{cheweihu,komathnk,zhuyan}@usc.edu,shri@sipi.usc.edu

ABSTRACT

The goal of this paper is to enable an objective understanding of gender portrayals in popular films and media through multimodal content analysis. An automated system for analyzing gender representation in terms of screen presence and speaking time is developed. First, we independently process the video and the audio content of a given movie, to estimate gender distribution of screen presence at shot level, and of speech at utterance level. A measure of the movie's excitement or intensity is computed using audio-visual features for every scene. This measure is used as a weighting function to combine the shot/utterance level gender distribution to compute gender representation for the entire movie. Detailed results and analyses are presented on seventeen full length Hollywood movies.

Keywords

Content analysis, gender representation, movie, multimodal.

1. INTRODUCTION

Automated analysis of multimedia content has become indispensable for organizing, indexing and navigating through vast amount of data. More recently, research in this area is being driven by the needs of facilitating and improving personal and social activities, insight generation, and interaction experience. Emphasizing the assistive, human-centric role that technology plays, this paper extends the application of multimedia content analysis to computational social science and media informatics. Specifically, our focus is on enabling an objective understanding of gender portrayals and biases in films. Thought leaders from the Hollywood movie industry, such as Geena Davis, have noted the presence of such inherent biases, and have both inspired research and championed steps for addressing them[1], including the work of this paper. Notably, this had led to research by social scientists and media experts who have repeatedly noted that women are highly underrepresented in popular films and media [12, 11]. Women comprise only 20-30% of all speaking characters in movies, and have much less screen presence and dialogs compared to their male counterparts [12]. To conduct such studies, researchers analyze large volume of media content *by hand*. This is extremely tedious, and calls for automated analysis. More importantly, automated content analysis can create opportunities for deeper, nuanced interpretation and analyses of multimodal content, which may not be even possible by manual inspection. For example, a recent work on automatic text analysis has reported significant differences between genders, and predictable gender patterns in terms of the language being used in movies [10]. .

In this paper, we propose to automatically quantify gender representation in movies from multimodal cues. Motivated by the aspects that media researchers and practitioners consider important, we automatically estimate *on-screen time* (from video) and *speaking time* (from audio) for male and female characters in movies. Since all parts of a movie are not equally important, it is reasonable to think that a character's presence (visual or audio) in the key parts of a movie is more important than that in less important parts. To capture this, we compute a measure of the movie's excitement (*scene intensity*) for every scene from video and music features. The raw on-screen and speaking time are then weighted by the corresponding scene intensity score to obtain a better understanding of gender representation in movies. We compute gender distribution on 17 popular Hollywood movies. Additionally, we perform various audio-visual analyses to study any inherent trends and biases in gender representation.

The objective of this work is twofold: (i) to provide a quantitative understanding of gender representation in media, and (ii) to build a scalable system that can scan large volume of data and automatically estimate quantities that are difficult to compute by hand.

2. PROPOSED APPROACH

Our gender representation framework has three main components: (i) on-screen time and (ii) speaking time estimation, and (iii) scene intensity computation.

2.1 On-Screen time estimation

Analyzing visual content in movies is extremely challenging, because the characters in movies exhibit significant variation in their appearances due to changes in scale, pose, illumination, expressions, camera angle etc. [1]. Our first task

[1] seejane.org

ICMI 2015, November 9–13, 2015, Seattle, WA, USA.

DOI: http://dx.doi.org/10.1145/2818346.2820778 .

Figure 1: Examples of success and failure of the face detector in movie frames.

Figure 2: Examples of confidence score for localizing landmark points on false and true faces.

Table 1: Shot-level face detection accuracy in %.

	Face detected	*No face detected*
Face exists	73.91	26.09
No face exists	6.82	93.18

Table 2: Comparison of manual and automatically computed on-screen time (hh:mm:ss).

		Manual	*Estimated*
On-screen time	M	01:15:30	01:13:16
	F	01:08:18	00:59:11

is to estimate the total time during which at least one female actor is present on screen, and the same when at least one male actor appears on screen. An obvious step is to detect and track all faces that appear on screen. However, popular face detectors, such as the Viola-Jones face detector [13], can reliably detect faces under limited conditions (e.g. frontal poses). Employing multipose face detectors [7] may increase accuracy, but they are not as efficient. Frame-to-frame tracking of multiple faces for long duration is also computationally expensive.

To mitigate the above issues, we cast screen time computation as a video shot labeling problem. A video shot is defined as a contiguously recorded sequence of frames. We assume that the same actors are present through out the duration of a shot. Hence, even if we can detect an actor only in a few frames within a shot, the actor's on-screen time includes the entire duration of the shot. Under this setting, computing gender-specific screen time requires detecting at least one face of each gender present in a shot. Our system next attempts to determine whether a given shot is a *male only* (`M`), *female only* (`F`) or a *male and female* (`MF`) shot by running a gender classifier on the detected faces. Our video processing framework comprises the following steps:

1) Partition into shots: First, we partition a given movie into N shots: $S_1, S_2, .., S_N$ with duration $t(1), t(2), ..., t(N)$ sec. We perform shot detection using an open source tool called the *ffmpeg*. This tool identifies the shot boundaries in a video stream by computing pixel-by-pixel differences between consecutive frames, and applying a threshold thereafter. Such pixel-based methods are robust to shot changes, and are popular in video processing literature [15]. We use a low threshold ($= 0.2$) value so that N is large (typically, 700 - 1500). Any S_i with $t(i) < 1$ sec is considered a false positive, and is merged with S_{i-1}.

2) Face detection: For each shot, a frontal face detector [13] is run on every frame. If at a frame, no face is detected, a profile face detector [13] is run on that frame (see Fig. 1 for examples of face detection). Our system is thus restricted to detect only the frontal and profile faces. Recall that we want to assign a label (`M`, `F` or `MF`) at shot level. Hence, the minimum requirement at this stage is to detect at least one human face of each gender present in the shot. If for a shot, no face is detected at all, it is labeled *empty*, and is not processed further.

3) False positives removal: Despite the high reliability of the face detector, false faces are detected. To remove false faces, a facial landmark point detection method [1] is used to localize 9 landmark points on eyes, nose and mouth corners (see Fig. 2) on each face. Originally developed for capturing mouth movement, this method returns a confidence score (c_v) as a measure of reliability of the localization. The idea is that in true faces, the landmark points will be localized with higher reliability, but for false faces c_v will be significantly low (see Fig. 2). If, for a face, c_v is greater than an empirically set threshold, it is considered to be a true face, and discarded otherwise.

4) Gender identification and Shot labeling: Our gender identification system is trained on the Labeled Faces in the Wild (LFW) database [5]. A subset of $6,000$ images (roughly equal male and female faces) are chosen from this database. Local Binary Pattern (LBP) [8] features (64 dimensional) are extracted from all faces. These features are used to train a binary Support Vector Machine (SVM) classifier with the radial basis function kernel. The overall gender classification accuracy on the LFW database is 90.85% (5-fold cross validation), with 99.42% accuracy for male, and 98.73% for female faces. This pretrained SVM is used to identify gender of the faces detected in movies. Consider n true faces extracted from a given shot S_i, where m faces are identified as male, and $n-m$ as female. A simple rule is used to label each shot: if $\frac{m}{n} \geq 0.9$, S_i is labeled `M`, i.e. $\ell_i \leftarrow$ `M` where ℓ_i is the label of S_i; for $\frac{n-m}{n} \geq 0.9$, $\ell_i \leftarrow$ `F`. Else, $\ell_i \leftarrow$ `MF`, implying the presence of both genders on screen. Each non-empty shot is thus labeled either `M`, `F` or `MF`. Total screen time for the female actors is $F_{scrn} = \sum_j t(j : \ell_j \in \{\texttt{F}, \texttt{MF}\})$. Similarly, the total screen time for the male actors is $M_{scrn} = \sum_j t(j : \ell_j \in \{\texttt{M}, \texttt{MF}\})$.

Validation: To evaluate the proposed system, we manually annotated 734 shots of duration ∼2 hours in total. An annotator labeled each shot as (`M`, `F` or `MF`), according to the presence of male and female actors. Shots with no recognizable human face (e.g. small faces or actors not facing the camera) are labeled *empty*. The performance of our video processing system is evaluated against the manual annotations at the *shot level*.

After running the face detector, false faces are removed by thresholding at $c_v \geq -20$. This step reduces the false positives from 15.9% to 6.8%. Face detection performance after false face removal is presented in Table 1. Gender identification accuracy is computed in terms of precision and recall. The average precision is 0.72 and average recall is 0.86. The error of misclassifying an `M` shot as an `F` shot is

Table 3: Acoustic gender identification results (in %) on 557 utterances.

		Computed labels	
		M	F
True labels	M	75.81	24.19
	F	12.17	87.83

Table 4: Comparison of manual and automatically computed speaking time (mm:ss).

		Manual	*Estimated*
Speaking time	M	17:09	14:18
	F	09:07	11:57

10.53%, and that of F as M is 13.77%. Overall screen time results are presented in Table 2.

2.2 Speaking time estimation

Our next task of computing speaking time of male and female characters in movies follows the steps below.

1) Speech region segmentation: Voice activity detection (VAD) is a crucial first step to automatic speech processing. Despite the availability of many algorithms, robust VAD in movie and web media content is extremely challenging due to highly noisy and non-stationary acoustic environment [3]. We use a VAD algorithm implemented in the OpenSMILE toolkit [2], where a long short-term memory recurrent neural network is trained, and tested on Hollywood films [3]. This algorithm is chosen because it exhibits high performance on movie data. The VAD is followed by a Bayesian information criterion (BIC)-based speaker segmentation to ensure each utterance is speaker-homogeneous.

2) Acoustic gender identification: Two key features in acoustic gender identification are *pitch* and the *Relative spectral transform - perceptual linear prediction* (RASTA-PLP) [14, 9]. Two Gaussian mixture models (GMM), pertaining to male and female speech, are trained on the TIMIT database [4], containing 6300 clean utterances from different speakers. Pitch, and 11-dimensional RASTA-PLP features are extracted, and used to train the GMM with 8 mixture components and a full shared covariance. The number of components and the type of covariance are chosen by cross validation. Using (randomly chosen) 75% of the samples as the training set, and the rest as the test set, we achieved 97.68% accuracy on female utterances, 98.12% on male utterances, with overall classification accuracy of 97.98%. The pretrained GMMs are then used to classify utterances from movies. To compute gender-specific speaking time, we assign a gender label (M or F) to each utterance. For each utterance, its likelihoods of being spoken by a male $P(\text{speech}|\texttt{M})$, and that by a female character $P(\text{speech}|\texttt{F})$, are computed. A gender label is assigned accordingly.

Validation: Our audio processing system is evaluated on a labeled database that we created. This database contains 557 utterances from the movie *Hope Springs*. An annotator manually assigned a label of M or F to each utterance. The gender identification system is then run on this database and results are compared against the manual annotations. Classification results in Table 3 indicates reasonable accuracy of our system on this challenging corpus. Table 4 presents overall comparison of manual and estimated speaking time.

2.3 Scene intensity and Gender representation

Scene intensity, in the context of this work, can be understood as a measure of excitement or activity in a movie scene. Measuring the excitement induced by a scene can be a challenging problem in itself. Motivated by the ideas like tempo or story intensity popularly used in content analysis [6], we compute a multimodal measure of scene intensity based on various low-level features.

For a given movie, we compute two visual features: *shot length* and *motion energy*. Following the procedure mentioned in section 2.1, we partition a movie into its constituent shots, and compute a feature vector $\mathcal{S}$ where the ith element is the number of frames in the ith shot. Motion energy at a video frame is computed as the total motion the frame has undergone with respect to its previous frame. Motion is estimated based on optical flow vector method. Motion energy feature $\mathcal{M}$ is computed at shot level by averaging across all frames in a shot. Since music plays a key role in communicating excitement to the audience, we extract a feature called *harmonicity* that detects the presence of music in a movie's audio stream. At a time instant, it is measured as the average periodicity in its neighborhood. The harmonicity feature $\mathcal{H}$ at shot level is computed by averaging its values over the entire duration of a shot. These three features, computed at shot level, are normalized to have zero mean and unit standard deviation. When combined linearly with equal weights [6], we obtain a measure of excitement at shot level. Let the intensity of the ith shot be denoted as $\mathcal{I}(i)^{(sh)}$. A scene in a movie is usually comprised a number of shots, and the scene boundaries can be detected using *ffmpeg*. Scene intensity for the kth scene $\mathcal{I}(k)^{(sc)}$ is computed as follows:

$$\mathcal{I}(k)^{(sc)} = \sum_{shot\ i \in scene\ k} \mathcal{I}(i)^{(sh)} \tag{1}$$

A high value of $\mathcal{I}(k)^{(sc)}$ indicates that a scene is more exciting, and is likely to be a key scene in a movie. Gender representation percentages are first computed at the scene level. Let the female representation (on-screen or speaking time) percentage for the jth scene be $\mathcal{G}(j)^{(f)(sc)}$. The scene intensity scores, $\mathcal{I}(k)^{(sc)}$, are then used to weight scene-level gender percentages to compute overall gender representation $\mathcal{G}^{(f)(sc)}$ as shown in Eq.(2).

$$\mathcal{G}^{(f)} = \sum_{scene\ j} \mathcal{I}(j)^{(sc)} \mathcal{G}(j)^{(f)(sc)} / \sum_{\forall scenes} \mathcal{I}^{(sc)} \tag{2}$$

3. RESULTS

Our audio and video processing systems are first run on the 17 full-length Hollywood movies independently, to estimate gender distributions of screen and speaking time. To explore any structure among movies on the basis of gender representation, a clustering attempt is made. Let F_{scrn}, F_{spk} denote the percentages of female screen and speaking time in a movie. Using these two measures, a feature vector $\mathbf{F} = [F_{scrn}, F_{spk}]$ is obtained for each movie. The movies are then subject to agglomerative hierarchical clustering, where the distance between two movies is the ℓ_2 distance between their features. The distance between two clusters (linkage criterion) is measured by weighted pair group with average. Applying a threshold at the 70% of the maximum distance, three clear clusters are observed (Fig. 3) indicating

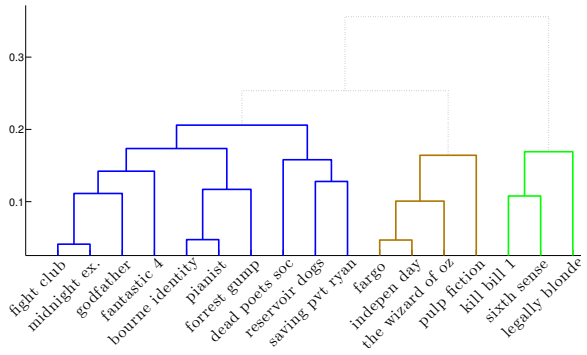


Figure 3: Agglomerative clustering of 17 movies on the basis of gender representation showing three clear clusters.

Table 5: Results of audiovisual co-occurance analysis normalized by speech (top)

Speech \ Face	None	M	F
None	0.26	0.49	0.24
M	0.20	0.51	0.28
F	0.16	0.50	0.33

similarity among the movies in terms of quantitative gender representation.

Audiovisual cooccurance analysis: A joint analysis of the audio and video streams is important to understand and compare gender representation across modalities. This joint analysis use the two output streams of gender labels obtained from the audio and video processing systems. We compute co-occurance scores (across all movies) of M/F appearance on screen with M/F speech (results in Table 5). A few interesting observations from Table 5: (i) in case of no speech, men are more likely to appear on-screen than women, (ii) when dialogs are spoken by women, male faces are still more likely to appear on-screen than females. Alternately, when a female actor appears on screen, male and female speech have equal probability.

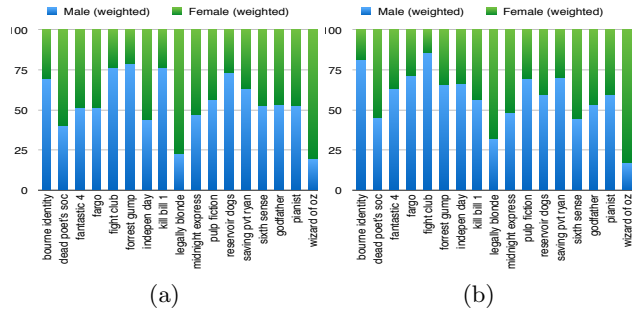
Next, we perform scene detection on each movie. Scene intensity is computed, and weighted average of the on-screen /speaking time are computed as quantitative measure of gender portrayal as described in section 2.3. Results in Fig. 4 suggests that *Bourne Identity, Fargo, Fight Club* and *Reservoir Dogs* have lower female representation compared to others. A summary of the gender representation results is presented in Table 6.

4. CONCLUSIONS

A multimodal framework for quantifying gender representation in movies is proposed. At this preliminary stage, our system computes simple quantities like on-screen and speaking time. These quantities are analyzed, and combined with scene intensity information to show how an objective understanding of gender representation in media can be developed. Our system is scalable, and can handle large number of movies.

Acknowledgment

The authors are grateful to Madeline Di Nonno (Geena Davis Institute on Gender in Media), Hartwig Adam (Google), Stacy Smith and Marc Choueiti (USC), and support from NSF and Google.org.



(a) (b)

Figure 4: Scene intensity-weighted gender distribution (in %) for (a) on-screen time, (b) speaking time.

Table 6: Summary of gender representation results

Measures	*Avg. female representation*
On-screen time	36.19%
Speaking time	41.06%
On-screen time (weighted)	45.12%
Speaking time (weighted)	42.17%

5. REFERENCES

[1] Arandjelovic, O., and Zisserman, A. Automatic face recognition for film character retrieval in feature-length films. In *CVPR* (2005), pp. 860–867.

[2] Eyben, F., Weninger, F., Gross, F., and Schuller, B. Recent developments in opensmile, the munich open-source multimedia feature extractor. In *ACM Mutimedia* (2013).

[3] Eyben, F., Weninger, F., Squartini, S., and Schuller, B. Real-life voice activity detection with lstm recurrent neural networks and an application to hollywood movies. In *ICASSP* (2013).

[4] Garofolo, J., Lamel, L., Fisher, W., Fiscus, J., Pallett, D. S., Dahlgren, N. L., and ZueE, V. Timit acoustic-phonetic continuous speech corpus. Linguistic Data Consort. Philadelphia.

[5] Gary, B. H., Manu, R., Tamara, B., and Erik, L. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Tech. Rep. 07-49, UMass, Amherst, Oct 2007.

[6] Guha, T., Kumar, N., Narayanan, S., and Smith, S. Computationally deconstructing movie narratives: an informatics approach. In *ICASSP* (2015).

[7] Li, S. Z., and Zhang, Z. Floatboost learning and statistical face detection. *IEEE Trans. PAMI 26*, 9 (2004), 1112–1123.

[8] Ojala, T., Pietikainen, M., and Maenpaa, T. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Trans. PAMI 24*, 7 (2002), 971–987.

[9] Parris, E., and Carey, M. Language independent gender identification. In *ICASSP* (1996), pp. 685–688.

[10] Ramakrishna, A., Malandrakis, N., Staruk, E., and Narayanan, S. S. A quantitative analysis of gender differences in movies using psycholinguistic normatives. In *EMNLP* (2015).

[11] Smith, S. L., and Choueiti, M. Gender disparity on screen and behind the camera in family films; the executive report.

[12] Smith, S. L., and Cook, C. A. Gender stereotypes: An analysis of popular films and tv.

[13] Viola, P., and Jones, M. Rapid object detection using a boosted cascade of simple features. In *CVPR* (2001), pp. I–511.

[14] Y. Zeng, Z. Wu, T. F., and Chan, W. Robust gmm based gender clas- sification using pitch and rasta-plp parameters of speech. In *Proc. ICMLC* (2006), p. 3376âĂŞ3379.

[15] Yuan, J., Wang, H., Xiao, L., Zheng, W., Li, J., Lin, F., and Zhang, B. A formal study of shot boundary detection. *IEEE Trans. CSVT 17*, 2 (2007), 168–186.